

ARMY RESEARCH LABORATORY



Chemical Compound Design Using Nuclear Charge Distributions

by B. Christopher Rinderspacher

ARL-TR-5935

March 2012

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory

Aberdeen Proving Ground, MD 21005

ARL-TR-5935

March 2012

Chemical Compound Design Using Nuclear Charge Distributions

**B. Christopher Rinderspacher
Weapons and Materials Research Directorate, ARL**

REPORT DOCUMENTATION PAGE

*Form Approved
OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

| | | | | | |
|--|-----------------------------|------------------------------|--|-------------------------------|---|
| 1. REPORT DATE (DD-MM-YYYY) | 2. REPORT TYPE | | 3. DATES COVERED (From - To) | | |
| March 2012 | | | | | |
| 4. TITLE AND SUBTITLE | | | 5a. CONTRACT NUMBER | | |
| Chemical Compound Design Using Nuclear Charge Distributions | | | | | |
| | | | 5b. GRANT NUMBER | | |
| | | | 5c. PROGRAM ELEMENT NUMBER | | |
| | | | 5d. PROJECT NUMBER | | |
| | | | 5e. TASK NUMBER | | |
| | | | 5f. WORK UNIT NUMBER | | |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) | | | 8. PERFORMING ORGANIZATION REPORT NUMBER | | |
| U.S. Army Research Laboratory ATTN: RDRL-WMM-G Aberdeen Proving Ground, MD 21005 | | | ARL-TR-5935 | | |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | | | 10. SPONSOR/MONITOR'S ACRONYM(S) | | |
| | | | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) | | |
| 12. DISTRIBUTION/AVAILABILITY STATEMENT | | | | | |
| Approved for public release; distribution unlimited. | | | | | |
| 13. SUPPLEMENTARY NOTES | | | | | |
| 14. ABSTRACT | | | | | |
| Finding optimal solutions to design problems in chemistry is hampered by the combinatorially large search space. We develop a general theoretical framework for finding chemical compounds with prescribed properties using nuclear charge distributions. The key is the reformulation of the design problem into an optimization problem on probability distributions in chemical space. In order to achieve tractability, a constrained search formalism on the nuclear charge distributions, which are non-negative, is used to reduce the dimensionality of the problem. Furthermore, we introduce approximations to the exact functional, as derived, for common design properties and constraints. | | | | | |
| 15. SUBJECT TERMS | | | | | |
| Chemical design, nuclear charge distribution | | | | | |
| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT UU | 18. NUMBER OF PAGES 20 | 19a. NAME OF RESPONSIBLE PERSON B. Christopher Rinderspacher |
| a. REPORT Unclassified | b. ABSTRACT Unclassified | c. THIS PAGE Unclassified | | | 19b. TELEPHONE NUMBER (Include area code) (410) 306-2811 |

Contents

| | |
|--|----|
| 1. Introduction | 1 |
| 2. Distributions as Optimization Variables | 1 |
| 3. Approximating \tilde{O} | 6 |
| 4. Enforcing Physicality | 7 |
| 5. Conclusions | 9 |
| References | 10 |
| Distribution | 13 |

INTENTIONALLY LEFT BLANK

1. Introduction

Molecular inverse design beyond the purview of drug design has of late enjoyed increasing popularity. Examples can be found in protein design (1–4) or high-hyperpolarizability materials (5–7). The design problem is complicated by the vastness of possible chemicals, termed chemical space. This space can be viewed as combinatorially complex, e.g., there are $20^8 \approx 2.6 \cdot 10^{10}$ octa-peptides of the naturally occurring amino acids alone. As a consequence, a variety of methods have been developed for discrete optimization in chemical subspaces (8–14). All these methods restrict the search to a predetermined subspace, often focused around molecular scaffolds. While this is convenient for reduced computational cost, it represents a bias, which precludes the discovery of unsuspected and truly novel solutions to the design problem, but also requires a priori knowledge to be successful. Instead of restricting the space explicitly, a “functional” description of the search space could remove the bias. Such a functional description must encode the target properties and constraints of the problem in a fashion that is amenable to computation.

The continuous optimization of chemicals used in the linear-combination-of-atomic-potentials (LCAP) method (15, 16) and the variation-of-particles density-functional-theoretical (VP-DFT) method introduces an important concept to overcoming the bias efficiently (17–19). LCAP and VP-DFT interpolate continuously between the Hamiltonians of various chemical species. Furthermore, recently an investigation into the reasons why chemical optimization is, relatively speaking, “easy” used probability distributions and expectations on the control variables to arrive at its conclusion (20, 21). It is this general idea that is extended and developed in the following, employing distributions of nuclear charges and associated potentials. Since the analytical form of such functionals is not generally known, we derive approximate functionals for the optimization of an electronic property for a molecule in the ground state (which is applied as a constraint), outlining a procedure for the development of other functionals.

2. Distributions as Optimization Variables

Our goal is to develop an optimization algorithm of chemical species based on non-negative distributions only. In order to be useful, it is required that the result can be identified (preferably uniquely) with a specific chemical. As a consequence, it is necessary to describe the properties of interest, the targets, and the constraints, as functionals in terms of these distributions. This is accomplished via probability distributions on chemical space. Furthermore, for properties with proper optima, the existence of and convergence on global

optima are proven for the exact functionals. We begin by defining rigorously the terms chemical space, conformational space, and configurational space as used later.

Definition 1. *The **compositional** space \mathcal{C} , i.e., the space of chemical compounds, is identified with the set of descendingly ordered finite sequences in \mathbb{N} . Each non-zero entry of the sequence denotes the charge of an atomic nucleus.*

For example, lithium hydride (LiH) corresponds to (3, 1) and beryllium hydride (BeH2) to (4, 1, 1). Due to the common restriction of charge, the charge is considered a separate property derivable from the composition and problem setting rather than an inherent component of the chemical definition, although it is possible to augment this picture without altering the general applicability of the following results. This choice of identifying the composition of chemicals reflects common usage in chemistry as well as providing a unique mathematical object for each such composition.

Definition 2. *The **configurational** space Γ is the set of finite sequences $((q_i, R_i))_{i \in I}$ in $\mathbb{N} \times \mathbb{R}^3$ of strictly descending lexicographical order ($i < j$ if $q_i > q_j$ or $q_i = q_j, R_i^\nu = R_j^\nu$ for $0 < \nu < \mu \leq 3$ and $R_i^\mu < R_j^\mu$). The coordinate system is chosen to conform to the rules:*

1. *The coordinate system is right-handed and comprised of principal axes through the center of mass.*
2. *If the coordinate system is ambiguous, then the axes are chosen to be a right-handed system such that the resultant sequence is minimal with respect to lexicographical order between sequences.*

The second entry in each couplet is the position of the respective atom in real space. Ambiguous coordinate systems may arise from high symmetry groups, which necessitates the second rule in defining the coordinate system. Now, let $\gamma : \Gamma \rightarrow \mathcal{C}$ map a configuration to its respective composition, i.e., $((q_i, R_i))_{i \in I} \mapsto (q_i)_{i \in I}$.

Definition 3. *The **conformational** space Ω_c of a composition c is the pre-image $\gamma^{-1}(c) \equiv \Omega_c \subset (\mathbb{R}^3)^{\ell(c)}$, where $\ell(c)$ is the length of the sequence c .*

For example, LiH corresponds to

$$\Omega_{\text{LiH}} = \{(x, y, z) \times (u, v, w) : x > u \vee (x = u \wedge y > v) \vee (x = u \wedge y = v \wedge z > w)\} \subset \mathbb{R}^3 \times \mathbb{R}^3.$$

Each of the preceding definitions refines the general notions of chemistry rigorously, which allows for well-defined treatment and analysis of chemical properties. We introduce the following terminology to simplify expressions.

- Let $C : \mathbb{N} \rightarrow \mathcal{C}$ be an enumeration of compositions.

- The nuclear charge of $C(i)$ is $\|C(i)\|_1 = \sum_{j=1}^{\ell(C(i))} (C(i))_j$, e.g., for LiH, the charge is 4.
- Let $P_i : \gamma^{-1}(C(i)) \equiv \Omega_{C(i)} \rightarrow \mathbb{R}$ be a probability distribution on the conformational space of composition $C(i)$, i.e., $P_i(x) \geq 0 \forall x \in \Omega_{C(i)}$ and $\int P_i d\mu(\Omega_{C(i)}) = 1$ with measure μ on $\Omega_{C(i)}$.
- Let $p' : \mathcal{C} \rightarrow \mathbb{R}$ be a probability distribution on compositional space, which is equivalent to $p : \mathbb{N} \rightarrow \mathbb{R}$, $p(i) = p'(C(i))$, i.e., $p(i) \geq 0 \forall i \in \mathbb{N}$ and $\sum_{i \in \mathbb{N}} p(i) = 1$.

For a probability distribution $\rho : \Gamma \rightarrow \mathbb{R}$, p and P_i can be defined such that the probability finding composition $C(i)$ at conformation $R_i \in \Omega_{C(i)}$ is $p(i)P_i(R_i)$, via $p(i) = \int_{\gamma^{-1}(C(i))} \rho(x) d\mu(\Gamma)$ and $P_i : \gamma^{-1}(C(i)) \rightarrow \mathbb{R}$, $x \mapsto \rho(x)/p(i)$.

With the help of these probability distributions, the following lemma now sets the stage for optimizations in chemical space.

Lemma 1. *For a bounded property $O : \Gamma \rightarrow \mathbb{R}$ that attains its global optimum exactly once, the optimal nuclear charge distribution is a distribution of discrete nuclei and charges.*

Proof. Let $O_{min} := O(X_{min})$. Then the expectation value for O is a linear functional of probability distributions, and $\langle O \rangle_\rho = \int_\Gamma \rho(x)O(x)dx$ has a minimal probability distribution $\rho_{min}(x) = \delta_{\Gamma,X_{min}}(x)$. Here $\delta_{\Gamma,X}$ is the Dirac distribution on Γ , i.e., the probability distribution on Γ such that $\delta_{\Gamma,X}(y) = 0 \forall y \in \Gamma \setminus \{X\}$. Hence, $\langle O \rangle_{\rho_{min}} = O(X_{min}) = O_{min}$ and the minimum is attainable. Due to $0 \leq |\int_\Gamma \rho(x)(O(x) - O_{min})dx| \leq \int_\Gamma |\rho(x)(O(x) - O_{min})|dx = \int_\Gamma \rho(x)(O(x) - O_{min})dx = \langle O \rangle_\rho - O_{min} = \langle O \rangle_\rho - \langle O \rangle_{\rho_{min}}$, there is no other distribution that produces a lower expectation value. Since $\langle O \rangle_\rho$ is linear and the space of probability distributions is convex, the minimum is global. Applying the argument to $-O$ yields the same for maxima. Since X_{min} is a single finite configuration, the associated nuclear charge distribution is just

$$q_{X_{min}}(R) = \sum_{j=1}^{\ell(X_{min})} q_j \delta^3(R - R_j), \quad (1)$$

where δ^3 is the Dirac distribution on \mathbb{R}^3 . \square

Analogously to equation 1, a nuclear charge distribution can be defined for any probability distribution ρ on Γ . For simplicity, assume that O_{min} is non-degenerate and, without loss of generality in the following, all compounds are considered to be neutral. We next clarify nuclear charge distributions.

Definition 4. The **nuclear charge distribution** $q(R)$ of a probability distribution ρ on Γ with p and P_i is given by

$$q_\rho(R) := \sum_i p(i) \int_{\substack{x=((q_j,R_j))_{j=1}^{\ell(C(i))} \in \gamma^{-1}(C(i))}} P_i(x) \sum_{j=1}^{\ell(C(i))} q_j \delta^3(R - R_j) d\mu(\Gamma).$$

The expectation charge of a distribution is $\bar{q} := \int_{\mathbb{R}^3} q(R) dR$. A non-negative distribution q is called **configurationally representable** if and only if there exists $x \in \Gamma$ such that $q = q_{\delta_{\Gamma,x}}$.

In order to perform optimizations over these distributions, it is helpful to understand the structure of the distributions derived from probability distributions on Γ . The convex space of potential nuclear charge distributions $\mathcal{Q} \subset \mathcal{L}_1(\mathbb{R}^3)$ is restricted to such distributions that are non-negative on \mathbb{R}^3 , whose coordinate systems coincide with the respective principal axes and centers of mass, and for which $\bar{q} \geq 1$. Since the space of probability distributions is convex and closed, so is the space of respective nuclear charge distributions $\mathcal{Q}_\Gamma \subset \mathcal{Q}$. The only nuclear charge distribution with $\bar{q} = 1$ is $q(R) = \delta(R)$, which corresponds to the hydrogen atom. Hence, the set of non-negative distributions can be convexly decomposed and the associated projection onto nuclear charge distributions is itself convex.

With this characterization it is possible to define unambiguously a functional on the space of non-negative distributions for a given bounded property.

Definition 5.

$$\tilde{O}[q] := \begin{cases} \min_{\rho: q_\rho = q} \{\langle O \rangle_\rho\}, & \exists \rho : q_\rho = q \\ \infty, & \text{otherwise} \end{cases}$$

This functional and its properties are our main objective and the remainder of this section is dedicated to characterizing this functional. The following lemma guarantees that the minima of \tilde{O} are global.

Lemma 2. \tilde{O} is convex, i.e., $\tilde{O}[\lambda q_1 + (1 - \lambda)q_2] \leq \lambda \tilde{O}[q_1] + (1 - \lambda) \tilde{O}[q_2]$.

$$\begin{aligned} \text{Proof. } \tilde{O}[\lambda q_1 + (1 - \lambda)q_2] &= \min_{\rho: \lambda q_1 + (1 - \lambda)q_2 = q_\rho} \langle O \rangle_\rho \leq \min_{\rho_1, \rho_2: q_1 = q_{\rho_1}, q_2 = q_{\rho_2}} \langle O \rangle_{\lambda \rho_1 + (1 - \lambda) \rho_2} = \\ &\min_{\rho_1, \rho_2: q_1 = q_{\rho_1}, q_2 = q_{\rho_2}} (\lambda \langle O \rangle_{\rho_1} + (1 - \lambda) \langle O \rangle_{\rho_2}) = \lambda \tilde{O}[q_1] + (1 - \lambda) \tilde{O}[q_2]. \end{aligned} \quad \square$$

Since $O_{min} = \langle O \rangle_{\rho_{min}} \leq \langle O \rangle_\rho$ for all probability distributions ρ , the global minima of \tilde{O} coincide with the minima of O , i.e., $\tilde{O}[q_{\rho_{min}}] = O_{min} \leq \tilde{O}[q]$ for all non-negative distributions q . Thus it is possible to arrive at the minimum of O by varying q and we have established the following important theorem.

Theorem 1. *For a bounded property O that attains its optimum, there exists a convex functional of non-negative distributions on \mathbb{R}^3 that optimizes to a global optimum of O .*

This result explains the successes of LCAP and VP-DFT as well as the discrete optimization approaches. It also generalizes the recent observation that optimization of molecular properties is considerably more successful than could be expected considering the vastness of the search space (20). Optimization in convex spaces is fairly efficient, and the reduction of the optimization space to distributions in only three dimensions instead of infinitely many is a considerable simplification.

Knowing the distribution without means of interpretation is not particularly useful. But if the global minimum of O is non-degenerate, then by lemma 1, q_{\min} corresponds to a single molecule or molecular system. This molecular system can be uniquely identified.

Corollary 1.1. *q_{opt} can be uniquely mapped to a composition and conformation.*

Proof. Let $q_{\text{opt}}(R) = \sum_i q_i \delta(R - R_i)$, then the configuration is the strictly descendingly ordered sequence of $((q_i, R_i))_i$. \square

In general applications, boundary conditions of physical viability need to be met. The introduction of such constraints does not violate the convexity arguments put forth earlier, since the subspace of probability distributions that are zero whenever the constraint is violated remains a convex set. Hence, we define the notion of physical distributions for common requirements.

Definition 6. *A distribution $P : \gamma^{-1}(c) \rightarrow \mathbb{R}$ for some composition $c = ((q_i, R_i))_i$ is considered **physical** if $\sum_i \left(\frac{\partial E(x)}{\partial R_i} \right)^2 = 0$ and $\prod_{i=1}^{\ell(c)} \eta_0(\det J_i) > 0$ for all $x \in \gamma^{-1}(c)$, where E denotes the energy and J_i is the i th principal minor of the Hessian of E with respect to the coordinates R_i , and η_0 is the Heaviside step-function with value 0 at 0. A distribution on Γ is considered **physical** if every P_i is physical. A distribution q on \mathbb{R}^3 is considered **physical** if $q = q_\rho$ for some physical distribution ρ on Γ .*

The energy as well as its derivatives are properties that are available over all of configurational space. Hence, they can be described as convex functionals of the charge distribution. Using Lagrange multipliers, it is therefore possible to incorporate physicality constraints using these respective functionals. Let $O : \Gamma \rightarrow \mathbb{R}$ be a property of interest for optimization bounded from below (above). In this case, the constrained functional becomes

$$\tilde{L} \equiv \langle L \rangle_\rho = \langle O \rangle_\rho + \lambda_g \langle G \rangle_\rho + \lambda_h \langle J \rangle_\rho, \quad (2)$$

where G is the gradient condition and J is the condition of positive-definiteness of the Hessian, with Lagrange multipliers λ_g and λ_h , respectively.

3. Approximating \tilde{O}

Since the exact functionals are in most cases unknown, it is necessary to approximate \tilde{O} . It is desirable that the exact functional can be systematically approximated with decreasing error. The following lemma proves useful in characterizing appropriate substitutes, as it obviates the need to be exact on anything but actually physical compounds.

Lemma 3. *If a functional W of charge distributions dominates \tilde{O} from above (below) and $W[q] = \min_{\{c \in \Gamma: q_c = q\}} O(c)$ for all configurationally representable distributions q , then W optimizes to the same nuclear charge distribution as \tilde{O} .*

Proof. $W[q] \geq \tilde{O}[q] \geq O_{min}$ (domination) and $W[q_{min}] = \tilde{O}[q_{min}] = O_{min}$, because q_{min} is a configurationally representable distribution. \square

We exemplify the use of lemma 3 in the development of an approximate operator to minimize a non-degenerate electronic property O , e.g., the inverse of the dipole moment, for neutral species in the ground state using only the nuclear charge distribution. For any configuration $c \in \Gamma$, the property $O(c)$ takes the value $\langle \psi | \hat{O} | \psi \rangle$, where \hat{O} is the corresponding operator on the space of wavefunctions $\mathbb{W}_{\|c\|_1}$ with $\|c\|_1$ electrons and $\psi \in \mathbb{W}_{\|c\|_1}$ is the normalized ground-state wavefunction of c . We define W such that it interpolates linearly between species with integer-valued numbers of electrons:

$$W[q] = \mu_1 \langle \psi_1 | \hat{O} | \psi_1 \rangle + \mu_2 \langle \psi_2 | \hat{O} | \psi_2 \rangle + \lambda_g G[q, \psi_1, \psi_2] + \lambda_h J[q, \psi_1, \psi_2] : \quad (3)$$

$$E \left[\frac{\lfloor \bar{q} \rfloor}{\bar{q}} q, \psi_1 \right] = \min_{\{\psi_3 \in \mathbb{W}_{\lfloor \bar{q} \rfloor}\}} E \left[\frac{\lfloor \bar{q} \rfloor}{\bar{q}} q, \psi_3 \right] \quad (4)$$

$$E \left[\frac{\lceil \bar{q} \rceil}{\bar{q}} q, \psi_2 \right] = \min_{\{\psi_4 \in \mathbb{W}_{\lceil \bar{q} \rceil}\}} E \left[\frac{\lceil \bar{q} \rceil}{\bar{q}} q, \psi_4 \right] \quad (5)$$

$$\mu_1 \lfloor \bar{q} \rfloor + \mu_2 \lceil \bar{q} \rceil = \bar{q}, \mu_1 + \mu_2 = 1 \quad (6)$$

The three constraints under consideration in this example are as follows:

1. The ground-state condition stated by equations 4 and 5 and energy functional E
2. The gradient condition G
3. The Hessian positivity condition J

We delay the discussion of the physicality conditions G and J to section 4 and focus on the general properties of W .

In order to apply lemma 3, it is necessary to establish that W dominates \tilde{O} and coincides with \tilde{O} on configurationally representable q . The ground-state condition obeys the conventional definition:

$$\hat{H}[q] = \sum_{i=1}^{\lceil \bar{q} \rceil} -\frac{1}{2} \nabla_i^2 - \int \frac{q(R)}{\|r_i - R\|} dR + \sum_{i < j} \frac{1}{r_{ij}}, \quad (7)$$

$$E_{nuc}[q] = \int \frac{Q(X, Y)}{\|X - Y\|} dXdY, \quad (8)$$

$$Q(X, Y) = \frac{1}{2} q(X) q(Y) (1 - \delta(Y - X)), \quad (9)$$

$$\underline{E}[q, \psi] = \left\langle \psi \left| H \left[\frac{\lfloor \bar{q} \rfloor}{\bar{q}} q \right] \right| \psi \right\rangle / \langle \psi | \psi \rangle + E_{nuc} \left[\frac{\lfloor \bar{q} \rfloor}{\bar{q}} q \right], \quad (10)$$

$$\overline{E}[q, \psi] = \left\langle \psi \left| H \left[\frac{\lceil \bar{q} \rceil}{\bar{q}} q \right] \right| \psi \right\rangle / \langle \psi | \psi \rangle + E_{nuc} \left[\frac{\lceil \bar{q} \rceil}{\bar{q}} q \right], \quad (11)$$

$$E[q, \psi_1, \psi_2] = \mu_1 \underline{E}[q, \psi_1] + \mu_2 \overline{E}[q, \psi_2], \quad (12)$$

where μ_1, μ_2 are as before. The Lagrange multipliers enforcing the gradient and Hessian conditions guarantee that unphysical q are mapped to infinity as they are for \tilde{O} . Hence, only the behavior of W on physical q is of interest. If q is physical and configurationally representable, $\hat{H}[q]$ and E_{nuc} are precisely the Hamiltonian and nuclear repulsion of the corresponding configuration c . Therefore, $W[q] = O(c)$ for physical, configurationally representable q , fulfilling the second requirement of lemma 3.

It remains to be shown that W dominates \tilde{O} for physical, not configurationally representable, distributions. The linear-interpolative construction of W with respect to \bar{q} reduces the discussion to arguments on integer-valued \bar{q} only. Without loss of generality, let q be an admixture of two distinguishable physical, configurations $x_1, x_2 \in \Gamma$ with $\|\gamma(x_{1/2})\|_1 = \bar{q}$ and ground-state wave functions $\psi_{x_{1/2}}$. Choose $\lambda \in [0, 1]$, then $q := \lambda q_{x_1} + (1 - \lambda) q_{x_2}$ and $\underline{E}[q_{x_i}, \psi_j] > \underline{E}[q_{x_i}, \psi_{x_i}]$, $i \in \{1, 2\}, j \neq i$, since ψ_j cannot be the ground state simultaneously for both distributions. Therefore, the gradient condition is violated and the Lagrange multiplier pushes the value of W to infinity. Otherwise, $W[q] = O(x_{1/2}) \geq \tilde{O}[q]$. Hence, the first requirement for lemma 3 is fulfilled as well. Since O_{min} coincides with a configurationally representable q_{min} , W attains O_{min} for q_{min} , and therefore, W converges upon minimization to the same composition and conformation as the general ensemble method does.

4. Enforcing Physicality

The gradient and Hessian conditions will be developed starting from a configurationally representable distribution. First, we look at the exact gradient functional G before

developing an approximate \tilde{G} . It is easily discernable that

$$G[q] := \min_{\rho: q_\rho = q} \int \rho(x) \sum_{j=1}^{\ell(x)} \|\nabla_j E_{gs}(x)\|^2 d\mu(\Gamma), \quad (13)$$

where we have used $E_{gs}(x)$ to denote the ground-state energy and ∇_j is the gradient of the j th nuclear position and is non-negative and zero only if ρ is physical. Given a configuration, the gradients are derived by perturbing infinitesimally the charge distribution at a specific point, i.e., $\delta(R - R_i) \frac{q_i}{\|X - R\|} \mapsto \delta(R - R_i - dR_i) \frac{q_i}{\|X - R\|}$. The gradient of the electron-nucleus energy is then taken with respect to the direction dR_i . Hence, the gradient is a small variation of the nuclear charge density. Therefore,

$$d^* \nabla_i E_{gs}(x) = \lim_{\epsilon \rightarrow 0} \int_{\mathbb{R}^3} \frac{\delta E_{gs}[q_{\delta_{\Gamma,x}}]}{\delta q_{\delta_{\Gamma,x}}}(Y) \frac{q_i(\delta(Y - R_i + \epsilon d) - \delta(Y - R_i))}{\epsilon} dY \quad (14)$$

$$= dq(R_i) \nabla \frac{\delta E_{gs}[q_{\delta_{\Gamma,x}}]}{\delta q_{\delta_{\Gamma,x}}}(R_i) = d^* \nabla q(R_i) \frac{\delta E_{gs}[q_{\delta_{\Gamma,x}}]}{\delta q_{\delta_{\Gamma,x}}}(R_i), \quad (15)$$

where d is a direction in \mathbb{R}^3 . Thus, we can rewrite equation 13 as

$$G[q] = \min_{\rho: q_\rho = q} \int \rho(x) \int \left\| q \nabla \frac{\delta E_{gs}[q_{\delta_{\Gamma,x}}]}{\delta q_{\delta_{\Gamma,x}}} \right\|^2 d\mu(\mathbb{R}^3) d\mu(\Gamma) \quad (16)$$

$$= \min_{\rho: q_\rho = q} \int \rho(x) \int \left\| \nabla q \frac{\delta E_{gs}[q_{\delta_{\Gamma,x}}]}{\delta q_{\delta_{\Gamma,x}}} \right\|^2 d\mu(\mathbb{R}^3) d\mu(\Gamma). \quad (17)$$

Two simple approximations given wavefunctions ψ_1/ψ_2 derive then by dropping all references to Γ and using the cumulative nuclear charge distribution,

$$\tilde{G}_1[q, \psi_1, \psi_2] = \int \left\| q \nabla \frac{\delta E[q, \psi_1, \psi_2]}{\delta q} \right\|^2 d\mu(\mathbb{R}^3), \quad (18)$$

$$\tilde{G}_2[q, \psi_1, \psi_2] = \int \left\| \nabla q \frac{\delta E[q, \psi_1, \psi_2]}{\delta q} \right\|^2 d\mu(\mathbb{R}^3), \quad (19)$$

where we use the linear interpolative form of the energy as used in equation 12. Since these expressions are non-negative and exact for configurationally representable nuclear charge distributions, $\tilde{G}_{1/2} = 0$ are good constraints. To satisfy lemma 3, the question remains whether $\tilde{G}_{1/2}$ may be zero for unphysical, non-configurationally representable nuclear charge distributions. We may restrict the discussion to integer-valued \bar{q} only. Therefore, the necessary conditions for fulfilling G_1 with the ground-state wave-function ϕ ,

$$\begin{aligned} G_1[q, \phi, \phi] &= 0 \\ \Rightarrow q(X) &= 0 \text{ or } \nabla \frac{\delta E[q, \phi]}{\delta q} = 0 \text{ a.e.,} \end{aligned} \quad (20)$$

may be considered under constraint of constant \bar{q} . We first show that $G_1 = 0$ is only given for countable sums of Dirac distributions. Then we show that these are composites of

physical distributions. If the set $N = \mathbb{R}^3 \setminus \{X \in \mathbb{R}^3 : q(X) = 0\}$ has non-zero measure, then $\frac{\delta E[q, \phi]}{\delta q}$ is a constant on the completion of N , which contradicts that $E[q, \psi, \psi]$ is a quadratic functional of q . Therefore, N must be a set of zero measure and, hence, q is a series of Dirac distributions.

As shown, G_1 is related to the norm of the first variation of the energy with respect to the nuclear charge distribution. If we assume without loss of generality that this distribution has integer charge, is not configurationally representable, and is a superposition of two distinct configurationally representable distributions, then the energy functional as a function of the convex parameter $\mu_1(\mu_2 = 1 - \mu_1)$, as defined before, is a concave functional. It follows that the gradient condition cannot simultaneously hold with the Hessian condition for such a distribution.

The Hessian condition J can be derived analogously to G . The Hessian analogue

$$J[q] = \min_{\rho: q_\rho = q} \int \rho(x) \int q_x(X) q_x(Y) S \left(\nabla_X \nabla_Y^* \frac{\delta^2 E[q, \psi, \psi]}{\delta^2 q} \right) dXdY d\mu(\Gamma) \quad (21)$$

$$J_1[q] = \int q(X) q(Y) S \left(\nabla_X \nabla_Y^* \frac{\delta^2 E[q, \psi, \psi]}{\delta^2 q} \right) dXdY \quad (22)$$

where $S(M) := \sum_{i=1}^3 \eta_0(-\det M_i)$, M_i is the i th minor of a 3×3 -matrix M , and η_0 is the Heaviside function, and is positive semi-definite for a configurationally representable energy minima. Extending finite-dimensional criteria to bilinear forms as in equation 21 and invoking the gradient condition renders the result that the Hessian matrix for the configurationally representable distribution q is only non-negative semi-definite if $J = 0$. We have thus demonstrated the equivalence of using G_1 and J_1 to the gradient and Hessian conditions for use in constrained optimizations on physical distributions.

5. Conclusions

We have laid out a general framework for the optimization of properties of chemical compounds using only nuclear charge distributions. The unknown exact functional is convex, which can be exploited for efficient optimization in discrete as well as continuous approaches. Furthermore, the convexity remains even for constrained problems. In absence of the exact functional, we have derived an approximate functional for the optimization of an electronic property under the constraint of physical viability. Ongoing efforts to develop general purpose electronic basis sets, e.g., wavelets (22–24) or finite element methods (25–27) open up the potential to use these same bases for efficient, systematic explorations of chemical space via nuclear charge distributions.

References

- [1] Bolon, D. N.; Mayo, S. L. Enzyme-like Proteins by Computational Design. *Proc. Nat. Acad. Sci. U.S.A.* **2001**, *98* (25), 14274–14279.
- [2] Havranek, J. J.; Harbury, P. B. Automated Design of Specificity in Molecular Recognition. *Nat. Struct. Biol.* **2003**, *10* (1), 45–52.
- [3] Park, S.; Xi, Y.; Saven, J. G. Advances in Computational Protein Design. *Curr. Opin. Struct. Biol.* **2004**, *14* (4), 487–494.
- [4] Mang, N. G.; Zeng, C. Reference Energy Extremal Optimization: A Stochastic Search Algorithm Applied to Computational Protein Design. *J. Comp. Chem.* **2008**, *29* (11), 1762–1771.
- [5] Dudiy, S. V.; Zunger, A. Searching for Alloy Configurations with Target Physical Properties: Impurity Design Via a Genetic Algorithm Inverse Band Structure Approach. *Phys. Rev. Lett.* **2006**, *97* (4).
- [6] Franceschetti, A.; Dudiy, S. V.; Barabash, S. V.; Zunger, A.; Xu, J.; van Schilfgaarde, M. First-principles Combinatorial Design of Transition Temperatures in Multicomponent Systems: The Case of Mn in GaAs. *Phys. Rev. Lett.* **2006**, *97* (4).
- [7] Piquini, P.; Graf, P. A.; Zunger, A. Band-gap Design of Quaternary (In,Ga)(As,Sb) Semiconductors Via the Inverse-Band-Structure Approach. *Phys. Rev. Lett.* **2008**, *100* (18).
- [8] Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220* (4598), 671–680.
- [9] Bohachevsky, I. O.; Johnson, M. E.; Stein, M. L. Generalized Simulated Annealing for Function Optimization. *Technometrics* **1986**, *28* (3), 209–217.
- [10] Muhlenbein, H.; Gorgesschleuter, M.; Kramer, O. Evolution Algorithms in Combinatorial Optimization. *Parallel Comput.* **1988**, *7* (1), 65–85.
- [11] Desmet, J.; Demaeeyer, M.; Hazes, B.; Lasters, I. The Dead-end Elimination Theorem and its Use in Protein Side-chain Positioning. *Nature* **1992**, *356* (6369), 539–542.
- [12] Allen, B. D.; Mayo, S. L. Dramatic Performance Enhancements for the Faster Optimization Algorithm. *J. Comp. Chem.* **2006**, *27* (10), 1071–1075.

[13] Keinan, S.; Paquette, W. D.; Skoko, J. J.; Beratan, D. N.; Yang, W. T.; Shinde, S.; Johnston, P. A.; Lazo, J. S.; Wipf, P. Computational Design, Synthesis and Biological Evaluation of Para-Quinone-Based Inhibitors for Redox Regulation of the Dual-specificity Phosphatase Cdc25B. *Organic & Biomolecular Chemistry* **2008**, *6* (18), 3256–3263.

[14] Rinderspacher, B. Christopher; Andzelm, Jan; Rawlett, Adam; Dougherty, Joseph; Beratan, David N.; Yang, Weitao. Discrete Optimization of Electronic Hyperpolarizabilities in a Chemical Subspace. *Journal of Chemical Theory and Computation* **2009**, *5* (12), 3321–3329.

[15] Wang, M. L.; Hu, X. Q.; Beratan, D. N.; Yang, W. T. Designing Molecules by Optimizing Potentials. *J. Am. Chem. Soc.* **2006**, *128* (10), 3228–3232.

[16] Keinan, S.; Hu, X. Q.; Beratan, D. N.; Yang, W. T. Designing Molecules with Optimal Properties Using the Linear Combination of Atomic Potentials Approach in an AM1 Semiempirical Framework. *J. Phys. Chem. A* **2007**, *111* (1), 176–181.

[17] von Lilienfeld, O. A.; Lins, R. D.; Rothlisberger, U. Variational Particle Number Approach for Rational Compound Design. *Phys. Rev. Lett.* **2005**, *95* (15).

[18] von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. Variational Optimization of Effective Atom Centered Potentials for Molecular Properties. *J. Chem. Phys.* **2005**, *122* (1).

[19] von Lilienfeld, O. A.; Tuckerman, M. E. Molecular Grand-canonical Ensemble Density Functional Theory and Exploration of Chemical Space. *J. Chem. Phys.* **2006**, *125* (15).

[20] Moore, Katharine W.; Pechen, Alexander; Feng, Xiao-Jiang; Dominy, Jason; Beltrani, Vincent J.; Rabitz, Herschel. Why is Chemical Synthesis and Property Optimization Easier than Expected? *Phys. Chem. Chem. Phys.* **2011**, *13* (21), 10048–10070.

[21] Green, David F. A Statistical Framework for Hierarchical Methods in Molecular Simulation and Design. *Journal of Chemical Theory and Computation* **May 2010**, *6* (5), 1682–1697.

[22] Harrison, R. J.; Fann, G. I.; Yanai, T.; Gan, Z.; Beylkin, G. Multiresolution Quantum Chemistry: Basic Theory and Initial Applications. *J. Chem. Phys.* **2004**, *121* (23), 11587–11598.

[23] Griffin, C. D.; Acevedo, R.; Massey, D. W.; Kinsey, J. L.; Johnson, B. R. Multimode Wavelet Basis Calculations Via the Molecular Self-Consistent-Field Plus Configuration-interaction Method. *J. Chem. Phys.* **2006**, *124* (13).

- [24] Massey, D. W.; Acevedo, R.; Johnson, B. R. Additions to the Class of Symmetric-antisymmetric Multiwavelets: Derivation and Use as Quantum Basis Functions. *J. Chem. Phys.* **2006**, *124* (1).
- [25] Pask, J. E.; Sterne, P. A. Finite Element Methods in Ab Initio Electronic Structure Calculations. *Modelling and Simulation in Materials Science and Engineering* **2005**, *13* (3), R71.
- [26] Gavini, V.; Bhattacharya, K.; Ortiz, M. Quasi-continuum Orbital-free Density-functional Theory: A Route to Multi-million Atom Non-periodic DFT Calculation. *J. Mech. Phys. Solids* **2007**, *55*, 697–718.
- [27] Suryanarayana, Phanish; Gavini, Vikram; Blesgen, Thomas; Bhattacharya, Kaushik; Ortiz, Michael Non-periodic Finite-element Formulation of Kohn-sham Density Functional Theory. *Journal of the Mechanics and Physics of Solids* **2010**, *58* (2), 256–280.

NO. OF
COPIES ORGANIZATION

1 ADMNSTR
ELEC DEFNS TECHL INFO CTR
ATTN DTIC OCP
8725 JOHN J KINGMAN RD STE 0944
FT BELVOIR VA 22060-6218

1 US ARMY RSRCH DEV AND ENGRG CMND
ARMAMENT RSRCH DEV & ENGRG CTR
ARMAMENT ENGRG & TECHNLGY CTR
ATTN AMSRD AAR AEF T J MATT
BLDG 305
ABERDEEN PROVING GROUND MD 21005-5001

1 US ARMY INFO SYS ENGRG CMND
ATTN AMSEL IE TD A RIVERA
FT HUACHUCA AZ 85613-5300

1 US GOVERNMENT PRINT OFF
DEPOSITORY RECEIVING SECTION
ATTN MAIL STOP IDAD J TATE
732 NORTH CAPITOL ST NW
WASHINGTON DC 20402

1 US ARMY RSRCH LAB
ATTN RDRL WMM G C RINDERSPACHER
BLDG 4600 RM C228
ABERDEEN PROVING GROUND MD 21005

3 US ARMY RSRCH LAB
ATTN IMNE ALC HRR MAIL & RECORDS MGMT
ATTN RDRL CIO LL TECHL LIB
ATTN RDRL CIO LT TECHL PUB
ADELPHI MD 20783-1197

INTENTIONALLY LEFT BLANK.